Discrepancies in Gender Codes

Tom W. Smith

NORC/University of Chicago

GSS Methodological Report No. 102

July, 2005

Introduction

While one most often thinks of error from sampling variability, the total-survey-error paradigm establishes that there are many different sources of error in surveys and that they may involve both random error (or variance) and systematic error (or bias) (Smith, 2005). Among the different sources of error are sample frame coverage, sampling variance, nonresponse, mode, interviewers, respondents, interviewer-respondent interactions (such as race-of-interviewer/race-of-respondent effects), question wording, question order/context, coding, data entry, analysis, documentation, and report writing/presentation.

A Case Study: Gender

This report discusses one example of error on a variable, the miscoding of gender. Gender is collected in two ways by interviewers. First, as part of the Household Enumeration Form (HEF) a roster of all household members is collected. This is used to determine who the randonly selected respondent will be through the use of a Kish, respondent-selection table and to collect information on household composition. Currently, the information collected for all regular household members (and certain visitors) are 1) their name, 2) relationship to householder (e.g. spouse, child, parent, etc.), 3) relationship to spouse/partner of householder (if applicable), 4) gender, 5) age, 6) marital status (if 13+), 7) whether staying elsewhere, and 8) reason for staying elsewhere (if applicable). In addition, by what lines people are listed on in the HEF, their status as regular household member or included visitor is indicated. This information is collected from a household informant. A household informant can be any knowledgeable household member (generally at least 14 years old) or, in rare instances, a non-household member.[1] Thus, the household information may or may not be provided by the person who is selected as the respondent. Second, gender is coded for the respondent by the interviewer usually at the end of the oral interview. Gender is recorded by observation. Thus, there are two measures of respondent's gender: 1) the HEF variables (GENDER1-14) supplied by the household informant and 2) the questionnaire variable (SEX) coded by the interviewer while observing the respondent.

To compare these two gender codings, one must use the variable RESPNUM to determine which of the household members is the respondent. Table 1 shows the % of GSS respondents on which these two indicators disagree. In general, disagreements were in the 3-6% range and average 4.5%. The one outlier was in 2002 when the GSS and the HEF were both done on CAPI. The respondent was selected by the CAPI program and RESPNUM was generated by the program, which, as we will see below, eliminated a major source of error. Also, unrelated to the coding of gender there was a more comprehensive review of the HEF variables that as a consequence further reduced miscodings of gender. This review also led us to decide that collecting household composition on a hardcopy HEF would both produce figures that more closely matched pre-2002 distributions and were more accurate overall (Smith and Kim, 2003).

---

[1]Non-household members who are informants make up 1% or less of all informants and mostly consist of non-residing caregivers and non-residing family members.

To examine the nature and reasons of the disagreements, we looked at a sample of cases from 2000 and 2004. The 2000 GSS was the last PAPI GSS and has both hardcopy HEFs and questionnaires. Knowing that HEF and questionnaire gender in the final data filed disagreed, we examined whether these codes were a) consistent with what was in the respective hardcopies and b) whether what was recorded in the hardcopies was consistent with information on respondent name, household relationship, and other information in the HEF and questionnaire. We examined a sample of 50 cases and in all cases it was clear what the respondent's actual gender was. This evaluation found that HEF gender was wrong for 48 cases and questionnaire gender was in error for 2 cases.

HEF gender was wrong in 22 cases due to the wrong respondent number being coded by the interviewer. This mistake was most commonly made because the interviewer wrongly used R's position in the Kish table rather than his/her position in the household listing. In 16 cases HEF gender was wrong due to the RESPNUM or gender value being data entered incorrectly. That is, the right value was in the hardcopy, but a wrong value was punched into the data set. In 4 cases HEF gender was in error because there was a relisting of household members due to the addition or deletion of a member or the change in the order of the listed members after the initial listing so that the members as eventually correctly listed and data entered did not agree with the originally correctly assigned respondent number. In 3 cases HEF gender was misrecorded in the hardcopy HEF and therefore subsequently in the data file. Finally, in 3 cases HEF gender was wrong for uncertain reasons. In two cases questionnaire gender was wrong. In both cases this was due to the mispunch of a correctly recorded gender. However, in cleaning cases over the years, instances of misrecorded gender apparently due to miscirclings have also been found. Using the figure that questionnaire gender (SEX) was wrong in 4% of the cases in disagreement, this means that an estimated 7-8 cases of the total of 2817 were miscoded on SEX or about 0.3% of the cases. Moreover, there was no evidence of bias in the errors on SEX since they represent random errors from mispunches and/or possibly miscirclings. A review of 2004 cases showed a similar pattern of most disagreements involving errors in the HEF gender of respondents rather than SEX and with no evidence of bias.

In most circumstances, the low level and random nature of errors in coding SEX would have no meaningful impact on data analysis. But it may have a small, but non-trivial, impact on the GSS measurement of sexual orientation. Sexual orientation is a behavioral rather than a psychological or self-identification measure in the GSS. The GSS asks "Have your sex partners in the last 12 months been... Exclusively male, Both male and female, Exclusively female?" (SEXSEX) and "Have your sex partners in the last five years been...Exclusively male, Both male and female, Exclusively female?" (SEXSEX5). By comparing SEX to these two measures, same- and opposite-gender, sex partnerships are determined (Smith, 2003). Since the proportion reporting same-gender, sexual partners during the last year is relatively small, it is much more likely that a heterosexual person would be wrongly classified as homosexual than the opposite misclassification. That is, a random miscoding of gender will increase the number of homosexuals and reduce the number of heterosexuals. In fact, for the two cases known to be wrongly coded on SEX in the sample of 2000 discrepancy cases, both are in turn wrongly classified as homosexual rather than heterosexual. With an estimate of 7.8 miscodes of SEX overall that might lead to a like number of misclassifications of sexual orientation (although taking people not responding to the sex items or having no sexual partners in the last year into

consideration reduces the likely number of such misidentifications to 4.9). Of course there is also a small possibility of some off-setting miscodes of homosexuals as heterosexuals. Since there were only 61 people identified as having a same-gender partner during the last year on the 2000 GSS, that might translate into around an 8.8% reduction in the estimated number of homosexuals (61 - 4.9/61). While not large, this impact is a) systematic, b) not trivial, and c) likely to attenuate relationships between sexual orientation and other variables (e.g. marital status).

To reduce analytical error due to miscodes of gender several strategies are possible. One possibility, such as adopted by Black, Gates, Sanders, and Taylor (2000) is to exclude from analysis any cases showing a discrepancy between gender from the HEF and questionnaire. This approach has the virtue of probably eliminating any cases on which gender is miscoded and, as a result, sexual orientation misidentifed. However, this is a conservative procedure since the vast majority of cases with discrepancies in fact have no miscodes of interviewer-assigned gender and thus no error is the reporting of sexual orientation. A second possibility is to conduct the analysis several ways to see how robust results are. This might include: 1) using SEX as coded, 2) excluding discrepant cases as Black et al. did, 3) using SEX checked against marital status and household composition, and 4) employing HEF gender instead of SEX (but clearly a less accurate approach overall).

Some miscodes in data sets are unavoidable. In PAPI surveys with separate data capture and data entry steps, there are two points at which errors may occur. Error at data-entry can be minimized by partial or complete double, data entry and verification. For CAPI surveys data capture and data entry are one step. For most variables in CAPI surveys there is no possible check once the interviewer enters a keystroke since there is neither any consistency checks that can be applied, nor any separate, original record to go back and validate the data against. Nor can one do double-entry, data verification since the data are captured and entered in one step. Gender is one of a handful of variables for which internal, consistency checks can be applied. But while such checks are routine for most GSS variables, they have not been applied to HEF and questionnaire variables like gender, in large part because for many years HEF variables were not processed and released as part of the analytical data set. Only after 1991 was it decided that HEF information was useful enough to be included. At that point the data were secured from earlier years and have subsequently been included in later years. The original questionnaires were not readily accessible for retrospective cleaning, but some internal and external consistency checks on the HEF data in subsequent years was established. For example, HEF information 1) has been checked for internal consistency on number of people listed, 2) has been used to fill in missing data in the questionnaire for variables like respondent's age (AGE) and the presence of a phone in the household (PHONE) and 3) has helped to resolve disagreements in the questionnaire on such variables as marital status (MARITAL) and number of earners (EARNERS). However, full consistency checks on other variables such as gender were not instituted.

The introduction of CAPI in 2002 enhances the ability to compare HEF variables with questionnaire variables while still in the field and thus to reconcile conflicting information. In addition, HEF variables have expanded in content over time and make up a larger share of the data than previously. For these reasons and to improve the quality of GSS data overall, cleaning will be expanded to run more consistency checks both internally within the HEF variables and between HEF variables and questionnaire variables such as gender, age, and marital status. This will lead to cleaner HEF data and will minimize miscodes of gender on SEX and, as a result,

reduce analytical misclassifications of sexual orientation.

References

Black, Dan; Gates, Gary; Sanders, Seth; and Taylor, Lowell, "Demographics of the Gay and Lesbian Population in the United States: Evidence from Available Systematic Data Sources," Demography, 37 (2000), 139-154.

Smith, Tom W., "American Sexual Behavior: Trends, Socio-Demographic Differences, and Risk Behavior," GSS Topical Report No. 25. Chicago: NORC, 2003.

Smith, Tom W., "Total Survey Error," in Encyclopedia of Social Measurement, edited by Kimberly Kempf-Leonard. New York: Academic Press, 2005.

Smith, Tom W. and Kim, Seokho, "A Review of CAPI-Effects on the 2002 General Social Survey," GSS Methodological Report No. 98. Chicago: NORC, 2003.